PROCEEDINGS OF TDWG

Biodiversity Information Science and Standards 1: e20533
doi: 10.3897/tdwgproceedings.1.20533

OPEN ACCESS

Conference Abstract

Quantifying quality: the "Apparent Quality Index", a measure of data quality for occurrence datasets

Francisco Pando ‡

‡ Real Jardin Botanico -CSIC, Madrid, Spain

Corresponding author: Francisco Pando (pando@rjb.csic.es)

Received: 23 Aug 2017 | Published: 23 Aug 2017

Citation: Pando F (2017) Quantifying quality: the "Apparent Quality Index", a measure of data quality for occurrence datasets. Proceedings of TDWG 1: e20533. https://doi.org/10.3897/tdwgproceedings.1.20533

Abstract

When making an initial assessment of a dataset originating from an unfamiliar source, a user typically relies on the visible properties of the dataset as a whole, such as, the title, the publisher, and the size of the dataset. Aspects of data quality are usually out of view, beyond some intuitions and hard to compare assertions. In 2007 at GBIF Spain we tried to correct that by developing an index that enables a user to assess the quality of Darwin Core datasets published by GBIF-Spain, and to track improvements in quality over time. Our goal was to create an index that is explicit, easy to understand, and easy to obtain. We dubbed that index "ICA" GBIF Spain (2010) for its name in Spanish "Índice de Calidad Aparente" (Apparent Quality Index). We say ICA measures "apparent quality", because, although unlikely, a dataset can have a high ICA, while its records are actually a poor reflection of the reality to which they refer. ICA summarizes data quality on the three primary dimensions of biodiversity data: taxonomic, geospatial and temporal.

In this contribution we will present the rationale behind the ICA, how it is calculated, how it works within the Darwin Test tool Ortega-Maqueda and Pando (2008), how it is integrated in the data publication processes of GBIF Spain, and some discussion and results about its utility and potential. We also compare ICA to the emerging framework for data quality assessmentTDWG Data Quality Interest Group (2016).

© Pando F. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Keywords

Data quality, biodiversity informatics, fitness for use, occurrence datasets

Presenting author

Francisco Pando

References

- GBIF Spain (2010) Índice de Calidad Aparente (ICA) / Apparent Quality Index (ICA). http://www.gbif.es/ICA.php. Accessed on: 2017-8-02.
- Ortega-Maqueda I, Pando F (2008) ARWIN TEST 3.3: Una aplicación para la validación y el chequeo de los datos en formato Darwin Core 1.2, Darwin Core 1.4 o Darwin Core Archive. http://www.gbif.es/Darwin_test/Darwin_test.php. Accessed on: 2017-8-02.
- TDWG Data Quality Interest Group (2016) A conceptual framework to enable DQ Assessment and DQ Management of biodiversity data in common and standardized way. https://tdwg.github.io/bdq/tg1/site/. Accessed on: 2017-8-10.